コンポーネント単位のWeb検索とそのランキング

吉田, 慎一郎
九州大学

殷, 成久
九州大学

中藤, 哲也
九州大学

廣川, 佐千男
九州大学

出版情報：情報処理学会研究報告. 2011, pp.1234-1237, 2011-03. 情報処理学会九州支部
バージョン：
権利関係：

情報処理学会研究報告
IPSJ SIG Technical Report

コンポーネント単位の Web 検索とそのランキング

吉田 慎一郎[†1]　殷　成 久[†1]
中 藤 哲 也[†1]　廣川 佐千男[†1]

詳細な検索エンジンを実現するには、Web ページの中から、必要な一部分 (コンポーネント) を選択的に抽出するラッパーが必要となる。本発表では、コンポーネントの重要度のランキングを用いることで、ラッパーを使わずに細粒 Web 検索エンジンを構築する手法を提案する。

Component-based Web Search Engine and Its Ranking

Shinichiro Yoshida,[†1] Chengjiu Yin,[†1]
Tetsuya NakatohQ and Sachio Hirokawa [†1]

A wrapper is a program that selectively extracts a necessary part (component) from Web pages. Automatic or semi-automatic wrapper construction is crucial to achieve a fine grained search engine for Web pages. However, this is not an easy task to achieve. This paper proposes a component-based search engine in which the content components gain a high score in the search results. Thus, the required segments for a query can be obtained without using a wrapper.

1. はじめに

Web 検索エンジンを作成するには、Web ページから情報を抽出することが必要である。しかし、1 つの Web ページにメインコンテンツ以外にも様々な情報が記載されている。例えば、サイト広告、サイトメニュー、関連記事などがある。その中で必要な情報はメインコンテンツだけである。メインコンテンツ以外の情報はノイズ　となり、検索の精度が低下するという問題がある。Web 検索エンジンの精度を向上するために、Web ページ分割アルゴリズムの研究が盛んである。粒度の小さい検索を実現するには、Web ページの中から、必要な一部分 (コンポーネント) を選択的に抽出するラッパーが必要となる。本発表では、コンポーネントの重要度のランキングを用いることで、細粒 Web 検索エンジンを構築する手法を提案する。

筆者らは、観光情報に対する新しい検索の仕組を検討している[8)–10)]。本研究で取り扱う分析情報として、まず、九州征服計画[⋆1]のサイト に掲載されてい 1303 件のブログ記事の HTML ファイルを収集した。それらの HTML ファイルから合計で 136,368 件のコンポーネントを抽出した。九州征服計画は九州観光推進機構[⋆2]によって評価認定された九州地域の観光関連のブログ記事を集めたものである (図 1)。それぞれの記事は、九州観光推進機構とは独立な外部サイトで公開されており、簡単な紹介とリンクから構成されている。

2. 半構造化文書に対するキーワード検索

ベクトルモデルに基づくキーワード文書検索と, XML に代表される半構造化文書検索の融合は, 検索技術についての研究として最もホットなテーマの一つである[6)]。一般の検索エンジンでも、単に検索結果の文書名のリストを表示だけではなく、ユーザーが与えた検索語を含む周辺部分 (スニペット) を表示している。論文[1)] では、XML 文書を対象として、キーワードに関連する部分木を抽出する方法を述べている。文献[2)] では、半構造文書の部分を検索するための演算法として定式化を試みている。文献[3)] では、不特定多数の構造を許容する半構造データベースの検索速度向上法について述べている。また、論文[4),5)] では、関係データベースにおけるキーワード検索の効率を分析している。論文[7)] では、XML 文書を分割することで、より粒度の小さい検索の効率化を提案している。

いずれの研究においても、半構造化文書の固定的な構造と、キーワードによる柔軟な検索を効率を下げることなくどのように統合するかが課題となっている。

ブログページも HTML で記述されており半構造化文書の典型ではある。しかし、利用目

†1 九州大学
Kyushu University
†2 大阪府立大学
Osaka Prefecture University
†3 大阪電気通信大学
Osaka Electro-Communication University

⋆1 http://www.welcomekyushu.jp/demon-kakka/
⋆2 http://www.welcomekyushu.jp

図 1 九州征服ブログ

的がはっきり決っている事務文書などと比べ、ブロガーの個人的意見や、企業広告などのように、そもそも目的が異なる情報が混在しているので、構造的な取扱いと重要度の評価は容易ではない。本稿では逆に、半構造化文書の最小構造を一つの単独文書と考えることで、通常の検索エンジンにより半構造化文書の検索を実現する。

3. コンポーネントのスコアによるページのスコア

ブログ記事のページには、ブログ作者が書いた記事本体の他に、広告や他の記事へのリンク、あるいは、読者からのコメントなども含まれている。目的によっては、広告やコメントも有意義の場合もあるが、ブログ検索のためには、それらは不要である。対象とするブログが一つのサイトにあり、共通のテンプレートで書かれていれば記事部分を抽出するラッパーで、必要な部分だけを抽出し、検索対象とできる。しかし、サイトが異ればテンプレートは異なる。最近では、同一サイトでも個人の好みによって独自のテンプレートを設定できるような場合もある。一方、記事本体を限定せずに HTML ページ全体を検索対象とすると、記事以外の部分の単語も特徴語として認識されることになる。その結果、検索結果の上位に表示されるページが、目的としている内容を含まないという問題が起る。

本論文では、HTML ページに対するインデックスとは別に、そのページに含まれる個々のコンポーネントに対するインデックスも構築し、二つのインデックスでランキングを行う方法を提案する。具体的には、HTML を木構造として表したときに葉として現れるテキスト部分を一つのコンポーネントとする。

3.1 アンカー・テキストの除外

本稿では、アンカー・タグの下に現れるアンカー・テキストはコンポーネントから除外する。アンカー・タグは外部へのリンクを表すものであり、そのページと関連はあるが、そのページの内容そのものではない。実際、個別ブログページの上下左右のフレーム部分に表示される項目のほとんどは、トップページや他の記事へのリンクや、広告へのリンクである。

3.2 コンポーネントのスコア

通常の文書と比べ、本稿で対象とするコンポーネントのサイズは非常に小さいものが多い。従って、一般的文書検索での特徴語評価指標である TF*IDF などは適さない。本稿では、そのコンポーネントに含まれる単語の異なり数と、HTML 木の根からの深さの積として、コンポーネントのスコアを定める。深さを使う理由は、「個別で詳細な内容ほど、HTML 構造の深い部分に置かれる」という考えるからである。それに対し、広告は他の記事へのリンクなどの、共通テンプレートを構成するコンポーネントは、比較的浅い部分に置かれていることが多い。また、個別記事では独自の内容が書かれるので、使われる単語の種類も多くなる。Web ページのスコアは、そのページに含まれるコンポーネントのスコアの総和として定義する。

4. 評価事例

1303 件のブログ記事に対して、コンポーネント単位の検索エンジンを GETA[⋆1] を使って実装した。GETA で提供される通常の文書スコア (weight) によるランクキングと、コンポーネント・スコアによるランクキングを比較するため、以下の 22 個のキーワードで実験を行った。

麺 スープ ラーメン 料理 ランチ チャーシュー 野菜 肉 豚 ネギ
酒 うどん グルメ 鶏 魚 醤油 卵 レストラン 食事 牛 米 とんこつ

図 2 は、それぞれのページに対する順位を 2 次元プロットしたものである。横軸が通常の文書としてのスコア (D スコア) による順位 (D 順位)、縦軸がコンポーネント・スコア (C スコア) に基く順位 (C 順位) である。表 2 はそのデータである。コンポーネント単位のラン

⋆1 http://geta.ex.nii.ac.jp

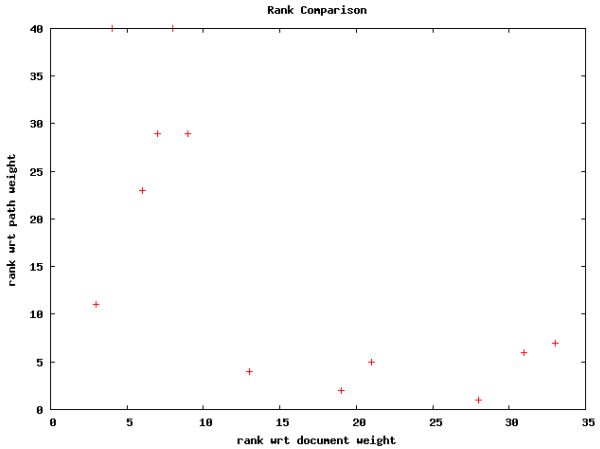

図 2 コンポーネントよる順位の変化

キングで順位が多きく変化したものを分析するために、D 順位、C 順位のどちらかで上位 10 位以内に入っているページに限定し、その順位差で整列している。これを見て分ることは、順位が下ったページは C スコアが小く、順位が上ったページは C スコアが大きい。表 2 は、具体的な各 Web ページに含まれるコンポーネントで入力したキーワードを含むものの内容を示している。順位が下ったページでは、短い文しか含んでないことが確認できる。順位が上ったページは、与えたキーワードを含み、詳細な内容が書かれている記事であることが分る。特に、順位が下ったページについて、入力したキーワードを含むコンポーネントはこの表に現れているものしかないことに注意する必要がある。すなわち、非常に短い文の中にキーワードが入っていただけなのに、そのページは検索結果の上位に表示されていたのである。また、逆に、通常のランキングでは上位 10 位に入らず表示されなかったページでも、与えたキーワードを含み、詳細な内容が書かれていれば、コンポーネントに基くランキングでは上位となり、検索結果として表示される。さらに、どのコンポーネントに含まれていたかも分るので、該当するコンポーネントを表示することができる。

5. まとめと今後の課題

多数の一般利用者が情報を発信する手段として、そして、形式張らない有用な情報を探す対象として、ブログは多くの利用方法が期待がされている。しかし、ブログ記事を含む

D 順位	C 順位	順位差	D スコア	C スコア	HTML
4	40	-36	0.0406	153	184.html
8	40	-32	0.0367	153	216.html
7	29	-22	0.0374	207	231.html
9	29	-20	0.0366	207	190.html
6	23	-17	0.0389	252	226.html
3	11	-8	0.0420	476	212.html
13	4	9	0.0348	880	541.html
21	5	16	0.0295	868	538.html
19	2	17	0.0309	1125	54.html
31	6	25	0.0260	830	939.html
33	7	26	0.0252	793	419.html
28	1	27	0.0272	1249	824.html

表 1 コンポーネント化によるスコアと順位の変化

変化	コンポーネント番号	内容
-	184-25	太麺の丸麺。
-	216-27	前回の麺三昧は
-	231-38	前回の麺三昧は。
-	190-25	ただ薄いか？と思いながら麺へ
-	226-22	前回の麺三昧は
-	212-81	モツ＆ちゃんぽん麺なんて最高ですね～（●＾o＾●）
＋	541-11	私はかなり久しぶりに来たんですが、ランチメニューを見ると、 日替わりのセット (800 円) に担々麺 (700 円)、酢豚や麻婆豆腐 （各 950 円）といった具合に、なかなか幅広い選択肢があるようです♪
＋	538-27	チャーシューも柔らかく旨みがあって美味しいんですが、この麺と 野菜のボリュームからするとやや寂しいので、トッピングの 「細切れ豚」(100 円) を頼むのが正解かもしれませんね
＋	54-63	とり天も冷麺も大好きな私！こりゃ参加せねばいかんでしょう
＋	939-119	野菜、魚介の旨みが凝縮されてミルクのようにコクのある クリーミーなスープは最高。スープに馴染んだ太麺の食感も最高。
＋	419-27	担々麺に、たっぷりのカボチャピューレと、素揚げした カボチャが入っています♪
＋	824-25	すると、麺に絡まるんです

表 2 順位が上るページと下るページ

Web ページは、情報提供手段として記事以外にも、様々な情報が一つの画面の中に埋め込まれており、必要な部分だけを抽出して利用することは困難である。また、ブログページの記述の仕方に共通規約はなく、多様な構造が使われている。本稿では、これらを固定的な構造として捉えるのでなく、逆に、Web ページを一番細かな単位まで分解し、コンポーネント単位の文書検索とすることで、必要な記事部分だけを抽出する方法を提案した。観光に関するブログに提案手法を適用し、通常の Web ページのランキングよりも有効であることを事例を通じて示した。

本稿では、タグで区切られる HTML 木構造の葉の部分をコンポーネントとした。ただし、ページそのものの情報ではなく、外部情報へのリンクであるアンカー・タグについては、アンカーテキストを除外している。コンポーネントの単位をどう決めるかは、様々な可能性が考えらえる。例えば、BR タグや P タグだけに限定する方法や、テキスト部分でも句点（。）で区切った文をコンポーネントとする方法なども、今後、検討する予定である。

参 考 文 献

1) 徳田隆志, 田島敬史, XML キーワード検索における解ノードの文脈の提示, C6-3,DEWS2008, 2008
2) 胡進, 清光英成, 大月一弘, 森下淳也, 半構造資料部分検索のための演算法の一考察 3C-i12, DEWS2006, 2006
3) 尾川広樹, 上土井陽子, 若林真一, 半構造データベースにおける検索速度向上のための Index Fabric 構造の拡張, DEIM Forum 2009 C7-6,2009
4) 張 麗茹, 大森 匡, 星 守, XML データ表現を考慮した関係データベースのキーワード検索, DEWS2008 C6-2, 2008
5) 本村徹太郎, 清水敏之, 吉川正俊, XML データに対するキーワード検索結果の分類とそれに基づく理解支援, DEIM Forum 2010 C7-3,2010
6) 絹谷弘子, 波多野賢治, 吉川正俊, 植村俊亮, キーワードを利用した XML 文書検索, 情報処理学会論文誌：データベース Vol.45, No.SIG 7 (TOD 22), pp.255-273, 2004
7) 波多野賢治, 絹谷弘子, 吉川正俊, 植村俊亮, キーワードを利用した XML 文書検索のための検索結果粒度決定法, DBSJ Letters Vol.2, No.1, 2003
8) X. Wu, S. Hirokawa, C. Yin, T. Nakatoh, Y. Tabata, Extraction and Comparison of Tourism Information on the Web, Proc. AROB2011, 2011
9) 殷 成久, 呉 小斌, 廣川 佐千男, 中藤 哲也, 観光イベントについての「といえば検索」の提案, 電子情報通信学会技術研究報告 110(301), pp.43-47, 2010
10) C. Yin, T. Nakatoh, S. Hirokawa, X. Wu, J. Zeng, A proposal of search engine “XYZ” for tourism events, Proc. JCAI2010, 2010